Data Analysis and Visualization in R (IN2339)

Exercise Session 5 - High dimensional visualization

Daniela Klaproth-Andrade, Felix Brechtmann, Julien Gagneur

Section 00 - Getting ready

1. Make sure you have already installed and loaded the following libraries:

```
library(ggplot2)
library(data.table)
library(magrittr)   # Needed for %>% operator
library(tidyr)
library(GGally)

library(pheatmap)
library(mclust)
```

Section 01 - Visualizing multiple variables

Gene expression measures the abundance of RNAs per gene. It is indicative of how active a gene is in a sample. Variations of gene expression across samples are indicative of the gene's role. The gene expression data in `cancer_data.rds` is a matrix of gene expression values of 20 genes across 30 tumor samples aimed at understanding the potential role of genes in cancer. .

Load the gene expression data in `cancer_data.rds` as a `data.table` with the following line of code:

```
expr <- readRDS("extdata/cancer_data.rds") %>% as.data.table(keep.rownames="tumor_type")
```

1. We are interested in the correlations between genes. Plot the pairwise correlations of the variables in the dataset. Which pair of genes has the highest correlation? *Hint:* remember that you can exclude a column `"colA"` from a data table `DT` with `DT[, -"colA"]`.

2. Visualize the raw data in a heatmap with `pheatmap`.

3. Does the latter plot suggest some erroneous entries? Could they have affected the correlations? Check by an appropriate plot the impact of these erroneous entries on the correlations. Substitute the erroneous values with missing values (NA) and redo the previous questions 2 and 3.

Section 02 - Heatmaps and Hierarchical clustering

1. Consider the full `iris` data set without the `Species` column for clustering. Create a pretty heatmap with the library `pheatmap` of the data without clustering.

2. Now, create a pretty heatmap using `complete linkage` clustering of the rows of the data set.

3. Annotate the rows of the heatmap with the `Species` column of the `iris` dataset. What do you observe when you compare the dendrogram and the species labels?

4. Obtain the dendogram of the row clustering using `complete linkage` clustering and partition the data into 3 clusters.

5. Create a pretty heatmap using `average` clustering of the rows annotated with the species and the complete linkage clustering results. What do you observe when you compare the dendrogram, the complete linkage results and the species labels?

6. Rartition the data into 3 clusters using the `average` clustering method.

7. Use the `table` function to compare the partitions from the complete and the average linkage clustering.

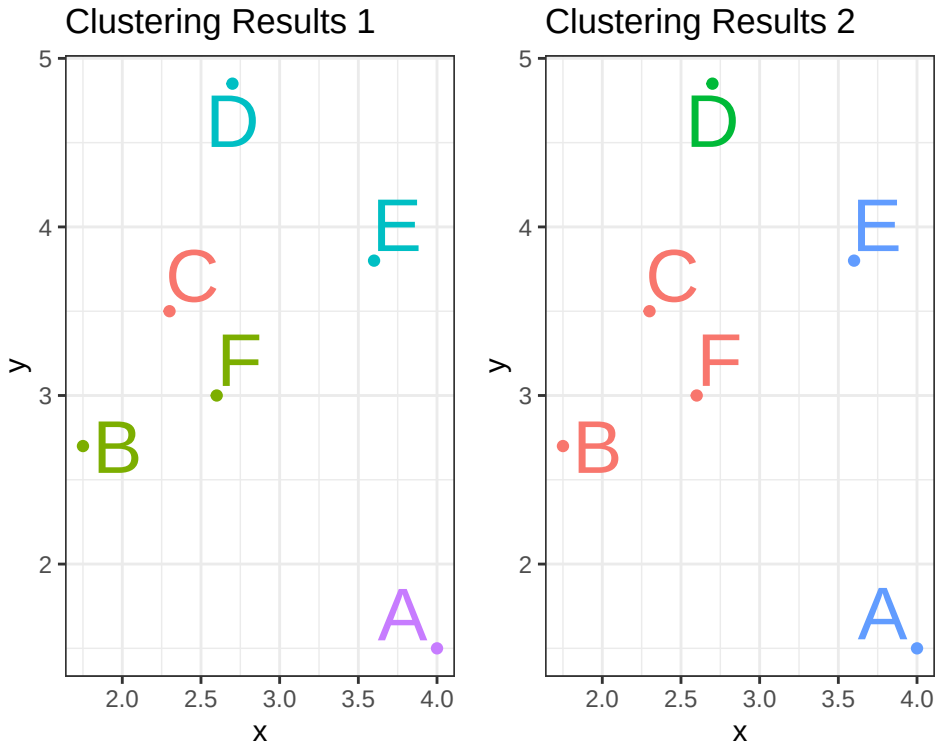
Section 03 - k-Means clustering

1. Perform k-means clustering on the iris data set with $k = 3$.

2. Create a pretty heatmap using `average` clustering of the rows annotated with the species, the hierarchical clustering results and the k-means results. What do you observe when you compare the dendrogram, the k-means results and the species labels?

Section 04 - Cluster comparison

1. Compute the Rand index between the two following clustering results from two different clustering algorithms.



2. Compute the Rand indices between the clustering results from the previous sections (complete, average and k-means) and species label. *Hint:* `rand.index()` from the library `fossil`.

3. [OPTIONAL] Visualize the pair wise Rand indices with a pretty heatmap. What is the best clustering in this scenario according to the computed Rand indices?

4. [OPTIONAL] Implement a function that computes the Rand index from two cluster label vectors (of same length). Verify the implemented function by comparing the rand indices computed before to the rand indices that can be computed with your function.

Section 05 - Dimensionality reduction with PCA

1. Let X be the `iris` data set without the `Species` column and only for the species setosa. Perform PCA on X. Make sure that you scale and center the data before performing PCA.

2. Which proportion of the variance is explained by each principle component?

3. Compute the projection of X from the PCA result and plot the projection on the first two principle components. *Hint:* `predict()` Additionally look at the biplot and come up with an interpretation of the first principal component.

4. Plot the first principal component against the other variables in the dataset and discuss whether this supports your previously stated interpretation. Discuss the interpretation in your Breakout Room.

5. Repeat the steps 1 - 4 for all species jointly (not only setosa). Discuss whether your original interpretation of the first principal component changed when performing the PCA for all species jointly. Use color to differentiate between the species in your plots.